DOCUMENT RESUME

ED 346 147 TM 018 482

AUTHOR Helgeson, Stanley L.
TITLE Assessment of Science Teaching and Learning Outcomes.
PUB DATE Apr 92
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Tests; *Computer Assisted Testing; Computer Managed Instruction; *Educational Assessment; Elementary Secondary Education; Evaluation Methods; Interactive Video; International Studies; National Surveys; Outcomes of Education; *Science Instruction; Scores; Technological Advancement
IDENTIFIERS Alternatives to Standardized Testing; Mandated Tests; National Assessment of Educational Progress

ABSTRACT

Indicators of current issues and concerns in the assessment of science teaching and learning outcomes are discussed. The greatest public attention to science test scores is devoted to scores from assessments that are external to schools, including national assessments such as the National Assessment of Educational Progress (NAEP), international assessments, and state-mandated assessments. Interest in attitudes toward science is an area drawing increased attention, although the field is hampered by a lack of a theoretical base. Another area of current interest is computer applications in assessment. Literature on computer assisted testing is reviewed, emphasizing: (1) the use of item banks; (2) evaluation of student data from laboratory exercises; (3) computer-generated homework problems and quizzes; (4) computerized feedback in the classroom; (5) advantages and drawbacks to computerized assessment; (6) computerized diagnostic testing; (7) computer managed instruction; and (8) the impact of interactive video. Alternative assessments have been the focus of several studies. The data collecting, storage, and analysis capabilities of the computer make alternative technology-based assessments increasingly attractive. There is a 70-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

STANLEY L. HELGESON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED346147

ASSESSMENT OF SCIENCE TEACHING AND LEARNING OUTCOMES

Stanley L. Helgeson
National Center for Science Teaching and Learning
The Ohio State University
Columbus, OH 43212

TM018482

BEST COPY AVAILABLE

ASSESSMENT OF SCIENCE TEACHING AND LEARNING OUTCOMES

Stanley L. Helgeson
National Center for Science Teaching and Learning
The Ohio State University
Columbus, OH 43212

The importance of assessment in science education is well recognized; thus, changes in assessment are also of significance. The goals for education have changed during recent years as have techniques of assessment. As noted by writers at the Bank Street College of Education: "Altering assessment practices is likely to affect curriculum, teaching methods, and students' understanding of the meaning of their work" ("Alternative Assessment and Technology," 1992:1). It is beyond the scope of this paper to consider all the ramifications of assessment and accompanying changes in science education. Rather, consideration will be given to a selection of topics as indicators of current issues and concerns.

Purposes of Assessment

Reasons for assessing student learning in science have been identified as: improvement of science instruction and programs; conveying expectations to students, parents, teachers, and administrators; monitoring the status of individuals, classes, districts, states, and the nation; and accountability (Raizen et al., 1989, p. 10). The areas of science to be assessed are generally agreed upon to include knowledge of facts and concepts, science process skills, science thinking and problem solving skills, skills needed to manipulate laboratory equipment, and the disposition to apply science knowledge and skills (Raizen et al., c.1991; Swain, 1985). The impact of assessment cannot be overlooked. Shavelson, Carey and Webb (1990) state that "Developers of scholastic tests have become overseers of a very powerful instrument of education policy making: achievement tests" (p. 692). Because of the intense interest in science achievement generated in recent

---

Panel presentation at the annual meeting of the American Educational Research Association, San Francisco, CA, April 20-24, 1992.

TM018482

years, achievement test scores have influenced the behavior of superintendents, principals, teachers, and parents. It is widely accepted that achievement tests drive the curriculum ("Alternative Assessment and Technology," 1992). The power of achievement test scores does not stop there; political decisions at the highest levels reflect concern about science scores as witness the fourth goal among six of the national education strategy presented by President Bush (AMERICA 2000: An Education Strategy).

The greatest public attention to science test scores is devoted to those resulting from tests that are external to the schools such as NAEP, Second International Science Study, and the various state-mandated tests. This area, therefore, shall be the starting point for the this consideration.

National and International Assessments

As has been widely reported, achievement scores on national assessments of science have not been encouraging. "Trends for 9-,13-, and 17-year-olds across five national science assessments conducted by NAEP from 1969 to 1986 reveal a pattern of initial declines followed by subsequent recovery at all three age groups. To date, however, the recoveries have not matched the declines" (Mullis and Jenkins, 1988:5). Because the 1990 NAEP assessment was scored in a manner different from that used with earlier assessments, direct comparison of results is not useful. However, a few of the findings are noted as indicators. Fewer than half the seniors demonstrated the knowledge and reasoning abilities needed to apply scientific knowledge to interpret data in tables and graphs, evaluate and design science experiments, nor did they possess in-depth knowledge of scientific information. Approximately two-thirds of the eighth graders and approximately one-third of the fourth graders were able to understand basic information in the physical sciences and basic ecological principles, and exhibited a beginning ability to interpret experimental results (National Center for Education Statistics, 1991:3). More than half the 17-year-olds tested in 1986 were deemed to be inadequately prepared to either perform competently jobs requiring technical skill or to benefit from specialized training; the 1990 results appear to offer only marginally better prospects (Mullis and Jenkins, 1988; NCES, 1990).

Differences in science proficiency between different racial and ethnic groups, although declining, still persist. Although no gender difference appears at grade 4 in the 1990 study, differences in science performance continue to appear at the higher grade levels and to increase with increase in grade level (Hueftle, Rakow and Welch, 1983; Mullis and Jenkins, 1988; NCES, 1990). Socioeconomic status also is related to science achievement; students from advantaged urban areas outperformed students from disadvantaged urban areas. Several factors in the home environment were also found to be related to science proficiency. At all three grade levels, students who had access to more reading materials performed better than students who had fewer materials available; students who watched six or more hours of television had lower average proficiency than students who watched less; and students who had both parents living at home had higher average proficiency than those from single parent families or those who lived apart from both (NCES, 1990:4).

By and large, these findings parallel the findings of international studies in which U.S. students took part. As a general rule, science achievement in the U.S. was lower than in other countries (Jacobson et al., 1987; Jacobson, Doran, Humrich and Kanis, 1988; Jacobson and Doran, 1988; Lapointe, Askew and Mead, 1992). It is, perhaps, interesting to note that the 1990 International Assessment of Educational Progress (IAEP) study found that, in nearly all of the 13-year-old populations, at least 10 percent of the students performed at least 15 points above the IAEP average and at least 10 percent performed at least 15 points below the average (Lapointe, Askew and Mead, 1992:12). Gender differences, much like those found in U.S. studies, appeared on a worldwide basis; only in biology did girls score higher than boys (Jacobson and Doran, 1988). However, in assessing science process laboratory skills, girls did about as well as boys, suggesting that teaching science via process tasks emphasizing manipulative activities may be a way to encourage girls to study science (Jacobson, Doran, Humrich and Kanis, 1988; Humrich, 1988; Kanis, Doran and Jacobson, 1990). In light of the relationship between television viewing and science proficiency indicated above, it is interesting to note that 13-year-olds in most countries are likely to spend much more time watching television (two to four hours per day) than doing homework (one hour or less per week in all countries except the Soviet Union) (Lapointe, Askew and Mead, 1992). The school factor most strongly correlated with achievement was the opportunity to learn (Jacobson, Doran, Humrich

and Kanis, 1988). In countries that had the highest science achievement scores there was relatively early specialization, and students studied the sciences for three or four years. (Jacobson and Doran, 1988).

In an effort to determine whether science and mathematics education has improved since the publication of A Nation At Risk, the Council of Chief State School Officers summarized data from national transcript studies, the NAEP data, and the State Indicators of Science and Mathematics (Blank and Engler, 1992). The report notes that students are now receiving more science instruction;from 1980 to 1987, 40 states increased science requirements for graduation. The rates of course enrollments by course level indicate that the proportion of students advancing through the secondary science curriculum has increased. "The percentage of students taking first year biology by the time they graduated increased from 75 percent in 1982 to 95 percent in 1990, the percentage taking chemistry went from 31 percent to 45 percent, and the physics enrollment went from 14 percent to 20 percent" (p. 3). Gender and race/ethnicity differences in course enrollments continue to exist. While science and math enrollments are below recommendations of the Excellence Commission, enrollments are increasing. States with higher requirements reflect more overall courses taken in science and mathematics and slightly more upper level courses taken (p. 4). Data from NAEP show slight average increases in science achievement from 1982 to 1990.

Summary Comments

Although the results of these studies show that science achievement is not at the level we might wish, there are some avenues that might be pursued. The disparity in scores between different racial and ethnic groups is decreasing, albeit not at the rate we might prefer. The evidence suggests that opportunity to learn is one of the strongest school factors in science achievement. Thus, altering the structure of the present pattern for science instruction would seem to promise greater success. Increasing the time students spend studying science by presenting the various disciplines over several years rather than in sequential years, as is the current practice, might do much to improve student success. Since there seems little likelihood of the schools bringing about major changes in students' home environment or socioeconomic factors, this appears to be a

change that the schools could initiate with some reason for optimism. Another factor that appears to have some promise is for increased emphasis on hands-on, manipulative approaches to instruction and to assessment. The evidence suggests that at least a portion of the existing gender difference in performance might be reduced by a science process orientation. Certainly, a large body of research evidence in science education supports such an approach, and has done so for many years. Finally, the summary data reported by the Council of Chief State School Officers indicates some gains over the past decade. Of particular promise is the fact that those states with the higher requirements have the higher enrollments and more upper level courses taken. Perhaps students respond to expectations established at the state level as well as at the classroom level.

State Mandated Assessments

With the call for accountability in the past decades came an increase in state-mandated assessment of student achievement. Most often this is accomplished by means of standardized tests. While testing is held by many to benefit education, the validity and value of traditional standardized testing is increasingly a subject of debate.

Because of the many negative factors attributed to the standardized tests used for much of student assessment, Herman and Golan (c.1992) conducted a study to discover more about the specific effects of standardized testing on teachers and classroom instruction. They surveyed more than 340 teachers in 48 different schools and identified definite relationships between standardized testing and the teaching and learning process. Among other things, they found that teachers felt strong pressure, especially from district administrators and the media, to improve their students' test scores. Administrators spent considerable time discussing with teachers ways to improve test scores and provided teachers with materials to support students' test taking skills. This finding is in agreement with results recently reported by Mertens (1992) in a study conducted in the state of New York.

Herman and Golan also reported that testing substantially influenced teachers' classroom planning. Teachers made sure that their instructional programs covered test objectives and many teachers looked at prior tests to assure a good match. Teachers also adjusted curricular scope and sequence based on test content and students'

prior performance. Further, teachers devoted substantial time to test preparation activities, test-wiseness instruction and practice tests p. 59-60). The results of the tests were, in the teachers' minds, of uncertain meaning and of uncertain value in school improvement. Teachers did not believe that standardized testing was helping schools to improve or that testing helped clarify school goals, provide useful feedback, or assess the most useful learning for students (p. 61-62).

The Elementary Science Program Evaluation Test (ESPET), administered to all fourth graders in New York State, was designed as an instrument to promote change. Cognizant of the danger that test scores alone can lead to unfortunate inferences, the ESPET developers added a non-mandatory component to the instrument to collect "information about student attitudes toward science, and about elements of the science program as perceived by students, administrators, teachers and parents/guardians...." (Mertens, 1992:2). These qualitative data were intended to help explain student performance on the required components of the test. Mertens chronicled the events which preceded and followed administration of the test in one relatively small, suburban district.

Mertens found, among other things, that the assumptions held by the superintendent, the principal, the teachers, and the science curriculum director were rarely clear to those involved and, instead of being shared, were, in fact, in direct conflict much of the time. The science director's attempts to improve the science program brought about changes that she and the teachers believed would improve the program, only to find that student scores on the ESPET declined. Shortly afterward, the test scores for every district in the county were published in the local paper. Resulting pressure from the administration to raise scores led to behaviors much like those described by Herman and Golan and away from the changes in the science program that had been sought by the teachers and science director. The dramatically improved science scores on the third year tests seemed to vindicate the actions taken to improve scores. Mertens states, "It is highly questionable, however, whether these higher scores can be interpreted to mean improvement in the school's science program" (p. 14). Rather than encouraging administrators to engage in a wide-ranging examination of the purpose, process, and structure of the program based on the test results, "the pressure to improve results puts educators in the position of having

to rely on 'quick-fix' activities that are narrowly focused on the tests themselves. This study clearly supports that conclusion" (p. 15).

Summary Comments

The negative attitudes of teachers toward standardized testing reported by Herman and Golan appears to be justified. Too often the tests appear to result in comparison of schools and districts, rather than in informed science program improvement. A documented outcome of standardized testing would seem to be the teaching of test-taking. In fact, there is some suggestion that testing may actually result in less science being taught, and that of less value. All too often program evaluations are reduced to single scores, with little evidence that these scores are related to the science program, instructional approach, or student understanding of science. On the brighter side is the fact that there are attempts to include other facets in statewide assessment. Florida, Missouri, California, Texas, and New York, to cite just a few examples, all are endeavoring to include some type of performance in the assessment effort, even if the attempts are not always as successful as one might like (Dana and Nichols, 1992; Clark, 1992; Comfort, 1992; Green, 1992; The State Department of Education, 1989). Finally, the advent of computer assisted assessment and performance based assessment offers the opportunity to evaluate a greater range of outcomes with greater flexibility.

Assessment of Attitudes in Science Education

Interest in attitudes toward science is by no means a new phenomenon, but it currently is drawing increased attention in the overall assessment efforts at national, state, and district levels. Several major reviews of studies related to student attitudes toward science have been conducted in the past two decades. Omerod and Duckworth (1975) summarized the results and implications of more than 500 attitude studies. Gardner (1975) evaluated results of studies and instruments used and noted that it was possible to distingusih two broad categories: attitudes toward science and scientific attitudes (p. 1). Gauld and Hukins (1980), in a reiview of scientific attitudes, identified as a major problem the lack of agreement about the meanings to be attributed to various terms that are used. Munby (1983a) examined the problems of assessment and instrumentation through review of more than 50 instruments used to as-

sess attitudes. Schibeci (1984) updated the research on attitude toward science and presented general conclusions and issues from more than 200 studies. Two problems appear consistently: lack of conceptual clarity in defining attitude toward science, and difficulty with instruments used to assess attitudes (Krynowsky, 1988). Drawing on the writing of Blosser (1984), Gardner (1975), and Munby (1983a), Germann (1988) states:

> First, the construct of attitude has been vague, inconsistent, and ambiguous. Second, research has often been conducted without a theoretical model of the relationship of attitude with other variables. Third, the attitude instruments themselves are judged to be immature and inadequate (p. 689)

Munby (1983b) used conceptual analysis to investigate the validity of the Scientific Attitude Inventory (Moore and Sutman, 1970), the most popular of the attitude instruments used in science education research. The SAI is a 60-item Likert-type instrument with a 4-point response scale, designed for high school students. Using conceptual perspectives developed in the study, Munby examined the instrument itself and 30 studies in which it had been used and concluded that the instrument's validity is highly questionable. He notes that the field of measuring attitudes is replete with instruments, but that these instruments are used in rather cavalier fashion, without heed to their reliability and validity (p. 157).

Based on the contention that validity testing of an instrument is dependent upon there being a conceptual link between the construct being measured and another construct, Munby, Kitto and Wilson (1976) describe the application of a multitrait-multimethod model to determine validity. This model involves first using two methods to measure the same trait. The correlation between the two measures should be significant since they are measuring the same trait. Then a second trait representing a construct coherent and converging to the first is identified. Two more instruments are then selected that measure this second trait with the intent that the scores of the first instruments should correlate significantly with the second pair of scores. Finally, it is necessary that the first two instruments are measures of a distinctive trait. That is, they must discriminate from other measures, a third pair, of a trait conceptually understood to have no relationship with the first two traits.

Following are two examples of attitude instrument development based upon both a theoretical foundation and empirical evidence. Krynowsky (1988) used the Azjen and Fishbein theory to guide the development of the Attitude Toward the Subject Science Scale (ATSS). This theory posits a theoretical relationship between attitude and behavior. This is interpreted to mean that "a student attitude toward the subject science is defined as a learned predisposition to an individual to respond in a consistently favorable or unfavorable way to performing behaviors related to the teaching/learning of the subject" (p. 579). Thus, an attitude assessment involves students evaluating the prospective performance of these behaviors. The ATSS was developed, refined, and tested for reliability and validity. Test-retest reliability coefficients of 0.82 and 0.84 were obtained. Validity was established by two techniques. First, teachers of two science classes were asked to rank students in terms of most positive attitude toward the subject science and the students' rank order was correlated with the students' rank order on ATSS scores. Spearman-rank order scores were 0.79 (n=25) and 0.65 (n=19) for the two classes. The second approach compared student ATSS scores to scores obtained on a reliable attitude toward the subject science scale, the School Science scale used in the British Columbia Assessment. The correlation of student scores was 0.70 (p. 581). It was the intent of the investigator that this example of using a theoretical/empirical basis for development could be further pursued by other science education researchers.

Drawing upon Fishbein's view that beliefs and behavioral intentions are determinants of attitude, Germann (1988) developed the Attitude Toward Science in School Assessment (ATSSA) as a measure of a unidimensional concept of attitude. The ATSSA "was to measure a single dimension of a general attitude toward science, specifically, how students feel toward science as a subject in school" (p. 694). The instrument consists of a Likert-type scale with five responses ranging from strongly agree to strongly disagree. The instrument was subjected to a series of pilot tests in which student scores were correlated with their own estimates of attitudes and with teachers' estimates, and to principal-component factor analysis. The revised instrument, consisting of 14 items, was then field tested in four studies. In all four studies, Cronbach's alpha estimates of reliability were all greater than 0.95. All 14 items were found to load on only one factor with consistent factor loadings in all four studies.

Discrimination was demonstrated by item-total correlations which ranged from 0.61 to 0.89. Attitude scores were compared from two classes, one of which was marked by poor discipline and inappropriate teacher behavior and the other by more experienced and skillful teaching. A t-test showed a significant difference between mean scores ($p = 0.001$) with the students in the more experienced class having more favorable attitudes. In a comparison of scores from two classes with equally skilled teachers a nonsignificant difference was found. While Germann notes that this evidence lacks experimental control, it speaks to the ability of the instrument to discriminate. In applying the instrument to study the relationship between attitude and achievement, he found relatively low correlations and concluded that other factors were more important in the kinds of learning measured by achievement tests. However, achievement including an evaluation of the consistency and quality of classwork as incorporated in a course grade seemed to be more strongly correlated to attitude.

Summary Comments

From the evidence it is clear that ambiguity of terms and quality of instruments are two serious problems facing those interested in assessing attitudes to science. The lack of a theoretical base has been identified in nearly all cases as a hindrance to assessment. Further, the lack of empirical support for most of the existing instruments exacerbates the situation. On the encouraging side, techniques for establishing the validity of an attitude assessment instrument have been detailed by Munby, Kitto and Wilson (1976). And, commendable attempts have been made to establish a theoretical foundation for assessment by Krynowsky (1988) and Germann (1988). Taken in combination, the guidelines and foundation exist, it now remains to capitalize on them. However, a note of caution is in order. In a meta-analysis of 49 studies on attitudes toward science, Haladyna and Shaugnessy (1982) reported that small differences exist in attitude between boys and girls; the effects of instructional programs on attitude are generally positive, but variable; and the teacher and the classroom environment play an important role in affecting attitude. As Germann points out, care must be taken in determining what factors are to be examined in relation to student attitudes.

Computer Applications in Assessment

In summarizing research findings on computer-based education, Waugh and Currier (1986) found that: (1) groups experiencing some kind of computer-based education attained test scores which were on average between .25 and .44 standard deviations higher than their comparison groups; (2) there was evidence favoring the use of computer-based education with academically disadvantaged students; (3) long term retention was no better for computer-based education than for other modes of instruction; (4) secondary students who experienced computer-based education had more positive attitudes toward computers than did their peers who did not experience computer-based education; and, (5) there was significantly less time required for computer-based education compared to conventional instruction. It should be noted that many of the studies summarized relied heavily on drill and practice modes of instruction. Such programs depend upon immediate feedback as a major function. While this may not fit the common perception of assessment, it is clear that it does in fact function in such a manner and that the immediate feedback may well have a positive impact on learning.

A common use of computers in assessment is to provide teachers with access to large banks of items for testing. These may range from specific topics such as medical biochemistry (Aesche and Parslow, 1988) for instructors of a given course, to a test bank designed for state assessment (Willis, 1988), to a broad range of juried test items which teachers anywhere in the country may access and download onto their own computers (Dawson, 1987). Once the item banks are in place, the computer may then be used to devise unique combinations of test items for each student and to use the results of those tests to develop remedial learning activities for each student. In each case, the computer can administer the quizzes, grade and record the results, and provide the student with immediate feedback (Dunkleberger, 1980). Use of the computer to

file test questions, assemble examinations, handle all records, produce and grade tests, and guide students to what should be done next enables testing to be done with an efficiency not possible from any teacher (Summers, 1984; Vogel, 1985; Heikkinen and Dunkleberger, 1985). Specially designed punch cards can be used for testing and grading large populations. The cards can be processed routinely at batch process stations by lab personnel who have little computer knowledge. This allows for cheap and easy marking and can be adapted to a wide variety of tests (Mihkelson et al., 1984).

Another form of formative assessment is the use of the computer to evaluate student data collected in laboratory exercises. Such checking of data and calculations is repetitive, prone to error, and not cost effective when done by humans. Computers, on the other hand, excel at this type of task (Harrison and Pitre, 1983, 1988). Programs used in this way are designed to check for realistic values, a range of data, and values clearly outside acceptable limits. When incorrect answers are given, students may be asked to redo their calculations and submit revised figures (May, Murray and Williams, 1985). The programs also may be designed to tentatively accept answers within a certain range, but to suggest that students return to places of potential error and check their work (Harrison and Pitre, 1988).

As part of a project to integrate computer-generated homework into physical science college courses, Milkent and Roth (1989) used computer-generated problems as homework assignments and monitored student progress with computer-generated multiple choice quizzes. They found that the use of the computer-generated homework significantly reduced the effectiveness of ACT scores as predictors of course achievement. Put in other words, as a result of the homework approach, students had greater opportunities for achieving mastery and for minimizing the potential influence of entry level aptitude and prior academic

preparation. This was in addition to the teacher advantages of an efficient system for homework management and freedom from bookkeeping procedures.

Incorporation of computers into science instruction often takes the form of microcomputer-based laboratories (MBL). Assessment is frequently a part of such a system. However, in some cases this means simply presenting multiple choice questions by means of the computer screen (Bross, 1986). If immediate feedback is not available, no learning gains may accrue to such computer use. Increased ease of data collection and processing may still make this approach to testing of value to the instructor. A more useful approach might be that described by Browning and Lehman (1988) for identifying student misconceptions in genetics problem solving. Four computer programs were presented and the students' responses were recorded and analyzed for evidence of misconceptions and difficulties in the problem solving process. Three main problem areas were identified: difficulties with computational skills, difficulties in the determination of gametes, and inappropriate application of previous learning to new problems. Evaluation of this type would seem to show considerable promise for remedial instruction and improved student learning.

Collins and Earle (1989-90) examined the effects of computer-based learning and computer-administered testing in an introductory biology class. They found that the greatest benefit was attained by those using the computer units in addition to attending regular lectures. Taking weekly computer-administered multiple choice tests also appeared to benefit students of middle and upper ability but not students of lower ability levels. That the use of weekly computer-tests can increase students' scores reinforces a finding of an earlier study (Collins, 1984). Although students benefitted from using either the computer learning units or the computer tests, the use of the two together did not result in even more gain, as might have been expected. Frequency of use of the

units appeared to be a factor in that the "frequent" user group achieved a much higher mean score and higher pass rate than did the "infrequent" user group.

The possibility that students were being disadvantaged by taking computer tests instead of written paper forms of the same tests was studied by Fletcher and Collins (1986-87). They found that students' mean scores on the computer-administered test and the written forms of the same test were roughly equivalent, and concluded that the students were not disadvantaged by taking the computer tests. The students indicated that most of them favored the computer-administered tests and cited several major advantages: (1) immediacy of scoring; (2) immediate feedback on incorrect answers; (3) more convenient, straight forward and easy-to-use; and (4) faster than written tests. Two major disadvantages were noted by the students: (1) not being to review all their responses at the end of the test and make changes; and, (2) not being able to skip questions and come back to answer them later (p. 42).

The converse case was studied by Jackson (1988) who attempted to discover whether a computer could give any significant educational advantage to the pupil. That is, could the computer improve pupil motivation during the test, by giving instant feedback and marking, thus improving understanding and hence give an enhanced score in a future test? (p. 809) The middle school science students who were tested by computer and given immediate feedback scored significantly higher in a later test using the same material than did those students who were tested using the traditional paper and pencil method. An additional gain for the teacher was the ability to conduct further analyses, such as test item analysis, on the computer-recorded student data; such analyses could not be easily carried out without computer administered testing.

The effects of microcomputer-administered diagnostic testing on both student achievement and attitudes were of concern to Waugh (1985). Students in one group were given the unit objectives and responded to a computer-administered diagnostic test consisting of one item per objective. The other group received the objectives and were assigned an out-of-class task of completing an objective specific mini-project. The results showed that microcomputer-administered diagnostic testing could positively influence the immediate achievement of students in science. Evidence did not, however, support the hypothesis that an exposure to diagnostic testing might influence continuing achievement. The findings indicated that the use of microcomputer-administered diagnostic testing was successful in increasing student achievement in science by an average of six percent with no loss of positive attitude toward school, learning, or science. The evidence further indicated that diagnostic testing might have played a role in arousing student interest in microcomputers.

Student attitudes were also the focus of a study by Knight and Dunkleberger (1977) in a comparison of computer-managed self-paced instruction with teacher-managed group-paced instruction for ninth grade students. The course consisted of large group lectures (31% of the overall time), small group seminars (46% of the time), and laboratory activities (31% of the time). The computer-managed self-paced group and the teacher-managed group-paced students received the same large group lectures and small group seminars. The computer group was allowed to self-pace through the laboratory activities while the teacher-managed group followed a group-pace. The computer served as an assessment and record keeping device for the computer-managed students. The quizzes were four-choice, multiple choice questions and students received immediate feedback after completing each item. Although the differing instructional approaches were applied only during the laboratory component of the course (31%), the positive reaction of the computer-managed

self-paced group was sufficiently strong to effect a significant difference in attitudes toward the study of science.

The impact of an emerging technology, interactive videodisc, was studied by Huang and Aloi (1991) in a first year biology course. The interactive video involved 17 menu driven chapters integrating computer text with laser disc images and computer graphics. The students were organized into groups with inter-group competition in answering true/false, multiple choice, and completion questions. The researchers compared, using an unpaired t-test, the proportion of students getting A, B, C, D, F, and W (withdraw) for 11 semesters prior to using interactive video with the proportions during the 5 semesters following its use. They found that the proportion receiving A's increased significantly ($p<.005$) following use of the interactive video. The percentage increases were: A's, 6% before and 18% after; B's, 21% before and 32% after; C's, 20% before and 36% after; D's 10% before and 4% after; F's did not change. Retention of students was also increased. The proportion of W's was 33% before interactive video use and 24% after. Thus, the use of interactive videodisc resulted in increased proportions of success at nearly all levels of achievement.

Interactive videodisc was also used as a tool in assessing science teachers' knowledge of safety regulations in school laboratories for purposes of teacher certification by the Connecticut State Department of Education (Lomask, Jacobson and Hafner, 1992). The program simulates a typical lab activity in a secondary school general science course and shows four student performing a simple lab experiment to identify unknown materials. The IVD assessment includes two stages: stage one deals with safety equipment and storage of chemicals and stage two deals with students' laboratory practices. The examinees are asked to assume the role of the lab teacher by viewing an interactive videodisc simulated classroom. The teachers are then asked to identify safety violations and to

suggest preventive or corrective measures. Subjects responses are recorded for later analysis and scoring (p. 1).

Summary Comments

There appear to be several advantages to incorporating some form of computer assistance in assessment. Immediate feedback to the students seems to be a consistent factor in increased achievement. Ease of test taking, together with improved record keeping, suggest improved efficiency for both students and teachers. The availability of large test item banks makes possible several intermediate quizzes with achievement gains apparently resulting. Such formative evaluation serves both as a diagnostic tool and as a remediation device, indicating where corrections are needed. The data collection capability of computer testing also permits more extensive data analysis, especially in the area of test item analysis, which in turn should yield more reliable and, presumably, more valid assessment. Two cautions must be noted, however. First, the simplicity of devising multiple choice, true/false, matching, and other objective tests can lull the teacher into simply doing a better job of assessing low level recall knowledge. Second, the linear nature of most computer testing does not allow the student to do back and reflect upon a particular item, nor to view the completed test as a whole to check for consistency of responses. The increased improvement and implementation of such emerging technologies as interactive video and hypermedia (Kumar, 1991) show high promise for overcoming both difficulties by providing opportunities for both improved levels of questions and increased flexibility in the testing process.

Alternative Assessments

Assessment attention primarily has been focused, perhaps understandably, on achievement in science. With regard to assessment of achievement, Shavelson, Carey and Webb (1990) note:

> Unfortunately, in an attempt to create achievement tests in science and other subjects that do not unduly favor one or another of the nearly infinite number of curricula in our country, the current technology produces tests that emphasize recall of facts and performance of

> isolated skills but tend not to measure students' conceptual understanding and problem-solving skills. Consequently, the current technology works against what many people value as education (p. 697).

Paper and pencil testing might suffice for determining how much, and what, students know if science programs were designed only to acquire content. As more problem-solving and process skills are incorporated into the science program, different forms of assessment are required to measure these learning outcomes (Meng and Doran, 1990). A variety of assessment methods are suggested by Meng and Doran. In addition to paper and pencil tests, there are practical tests, observations, discussions (or interviews), practical tests in which students manipulate materials, and projects and written work in which the students can demonstrate investigative research or construct something. It is necessary, of course, that the type of assessment be appropriate for the learning outcomes being measured. Although paper and pencil tests are currently regarded with suspicion, they are adequate for measuring knowledge of certain kinds of content, such as facts and terminology. They do not, however, effectively measure whether or not students can apply their knowledge. Application of knowledge involving problem-solving and process skills are better assessed by means of practical tests or observations of student performance.

In the belief that assessment techniques beyond the traditional paper and pencil tests were needed, Doran, Boorman, Chan and Hejaily (1992a) conducted a study to develop and validate instruments to assess the level of laboratory skills possessed by students completing the high school science courses (biology, chemistry, and physics). More than 1000 students from 35 schools participated in the study. It was decided that the "whole investigation" format of laboratory practical testing would be the model used and, in each of the science areas, tests were developed around six laboratory tasks (Doran, Boorman, Chan and Hejaily, 1992b). The model used was one developed by Tamir, Lunetta and colleagues and is composed of stages that are congruent with the prelab/lab/postlab format of many inquiry-oriented science programs (Lunetta, Hofstein and Giddings, 1981; Uri and Hofstein, 1982). The three stages used were planning, performing, and reasoning. A two-part test was designed to account for the different kinds of skills needed in the planning/design stage and those required in the other two stages. The

scoring system developed was general in the sense that it was applicable across each of the tasks in biology, chemistry, and physics. The scoring was subjective in that the test booklets were scored by raters. Thus, test reliability, inter-rater agreement, and correlation between raters had to be determined. Analyses of the data indicated that the reliabilities and correlations were sufficient to warrant further development and investigation.

A national assessment program in the United Kingdom is described by Burstall (1986). The target audience for science was 11-, 13-, and 15-year-olds. The overall assessment included both an oral and a practical component. The oral test was designed to determine the student's ability to communicate effectively and deals with general rather than science-specific information. The science practical test included both a script to guide the examiner's verbal assessment of student understanding and a checklist to record observations of the student's performance. The major advantages of the practical assessment were found to include: (1) All question are oral; poor readers are not penalized. (2) Students may ask for clarification of the test. (3) The pace and extent of the testing can be adjusted to suit the student. (4) Students have the opportunity to retract or amend an answer. (5) Assessors are permitted to prompt in order to direct students toward an appropriate strategy. (6) Assessors are able to observe the method and problem-solving strategies of the student (p. 18).

Keys (1992) describes a study to examine the converse of the procedure described above in that the students' written laboratory reports were analyzed for evidence of conceptual and procedural understandings in science. The eighth grade students involved did two laboratory activities, one on the inclined plane and one on the action of levers. The students had been previously trained to write laboratory reports in full sentences using a specified structure including, problem, hypothesis, materials, procedure, data/observations, conclusion and discussion sections (p. 4). Questioning prompts developed by the researcher were included in the report-writing stage to encourage the students to write more about the procedures and the concept involved in the activities. The laboratory reports were analyzed for conceptual and procedural understandings, and a scoring guide was developed using propositional analysis and rating scales. No reliability was determined for the scoring guide, this is intended to be further investigated and refined.

Qualitative analyses of the students' reports revealed many naive conceptions and instances where students used the data they had collected to support their naive conceptions. There was a tendency for students who had collected reasonable data to draw conclusion which the data did not support. In some cases, students drew conclusions in the opposite direction from that which could be inferred from the data. Others had unreasonable data, but still arrived at correct conclusions. Another tendency was to oversimplify the results in an attempt to make the drawing of conclusions more manageable. While there are still some evident problems to be solved with this approach to assessment, it does provide another option to be further pursued.

Many other options exist, of course. The March, 1992, issue of Science Scope is devoted to alternative forms of assessment. Similarly, a chapter in Science Assessment in the Service of Reform also includes descriptions of assessment alternatives (Kulm and Malcom, 1991). Articles are included on performance-based assessment, portfolio assessment, group assessment (involving team approaches), concept mapping, scoring rubrics (techniques), dynamic assessment, and assessment for individual differences. While these approaches, for the most part, represent still-emerging options with validity and reliability still to be established, they nonetheless offer ways to broaden assessment and improve the prospects for gaining a more complete picture of student understanding.

Summary Comments

Alternative approaches to assessment offer both encouragement and opportunities. The encouragement stems from the possibilities we gain to develop a more complete perspective of student knowledge and understanding. The opportunities derive from the work still to be done in determining the validity and reliability for most of the approaches. Most of these alternatives are more time consuming than is traditional testing. With the increased availability and flexibility of computer and related technology, the cost in time and effort can be sharply reduced. The additional data collecting, storing and analysis capability made possible with the computer makes alternative forms of assessment increasingly attractive.

REFERENCES

Aesche, Darryl W. and Graham R. Parslow. (1988). "Use of Questions from the Medical Biochemistry Question Bank with the 'Q' Instruction Package." Biochemical Education, 16(1): 24-28, January.

"Alternative Assessment and Technology." (1992). News from the Center for Children and Technology. 1(3): 1-5, March. Bank Street College of Education, NY.

"AMERICA 2000: An Education Strategy." (1991). Washington, DC: U.S. Department of Education.

Blank, Rolf K. and Pamela Engler. (1992). "Has Science and Mathematics Education Improved Since A Nation at Risk?" Science and Mathematics Indicators Project, Washington, DC: Council of Chief State School Officers, January.

Blosser, Patricia E. (1984). "Attitude Research in Science Education." Information Bulletin No. 1. Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education, The Ohio State University.

Bross, Thomas R. (1986). "The Microcomputer-Based Science Laboratory." Journal of Computers in Mathematics and Science Teaching, 5(3): 16-18, Spring.

Browning, Mark E. and James D. Lehman. (1988). "Identification of Student Misconceptions in Genetics Problem Solving via Computer Program." Journal of Research in Science Teaching, 25(9): 747-761, December.

Burstall, Clare. (1986). "Innovative Forms of Assessment: A United Kingdom Approach." Educational Measurement: Issues and Practice, 5(1): 17-22, Spring.

Clark, Dave. (1992). "Missouri's Process Skills Approach." Science Scope, 15(6): 56, March.

Collins, Michael A.J. (1984). "Improved Learning With Computerized Tests." The American Biology Teacher, 46(3): 188-190.

Collins, M.A.J. and P. Earle. (1989-90). "A Comparison of the Effects of Computer-Administered Testing, Computer-Based Learning and Lecture Approaches on Learning in an Introductory Biology Course." Journal of Computers in Mathematics and Science Teaching, 9(2): 77-84, Winter.

Comfort, Kathleen B. (1992). "Crime Solvers in California." Science Scope, 15(6): 56-57, March.

Dana, Thomas M. and Sharon E. Nichols. (1992). "Assessing the State of Science Education Reform in Florida." Paper presented at the annual meeting of the National Association for Research in Science Teaching, Boston, MA, March 21-25.

Dawson, George. (1987). "10,000 Science Questions by Modem." The Science Teacher, 54(3): 41-44, March.

Doran, Rodney L., Joan Boorman, Fred Chan and Nicholas Hejaily. (1992a). "Alternative Assessment of High School Laboratory Skills." Paper presented at the annual meeting of the National Association for Research in Science Teaching, Boston, MA, March 21-25.

Doran, Rodney L., Joan Boorman, Alfred Chan and Nicholas Hejaily. (1992b). "Successful Laboratory Assessment." The Science Teacher, 59(4): 22-27, April.

Doran, Rodney L. and Willard J. Jacobson. (1991). "Science Achievement in the United States and Sixteen Other Countries." In S.K. Majunder, et al. (Eds.), Science Education in the United States: Issues, Crises, and Priorities. Easton, PA: Pennsylvania Academy of Sciences.

Dunkleberger, Gary E. (1980). "Applying 20/20 Hindsight to Self-Pacing." The Science Teacher. 47(8): 32-34, November.

Fletcher, P. and M.A.J. Collins. (1986-87). "Computer-Administered Versus Written Tests--Advantages and Disadvantages." Journal of Computers in Mathematics and Science Teaching, 6(2): 38-43, Winter.

Ganiel, Uri and Avi Hofstein. (1982). "Objective and Continuous Assessment of Student Performance in the Physics Laboratory." Science Education, 66(4): 581-591.

Gardner, P.L. (1975). "Attitudes to Science: A Review." Studies in Science Education, 2: 1-41.

Gauld, C.F. and A.A. Hukins. (1980). "Scientific Attitudes: A Review." Studies in Science Education, 7: 129-161.

Germann, Paul J. (1988). "Development of the Attitude Toward Science in School Assessment and its Use to Investigate the Relationship Between Science Achievement and Attitude Toward Science in School." Journal of Research in Science Teaching, 25(8): 689-703, November.

Green, Barbara. (1992). "Statewide Assessment in Texas." Science Scope, 15(6): 56-57, March.

Haladyna, Tom and Joan Shaughnessy. (1982). "Attitudes Toward Science: A Review." Science Education, 66(4): 547-563.

Harrison, David and John M. Pitre. (1983). "Computerized Lab Test on Error Analysis." The Physics Teacher, 21(9): 588-589, December.

Harrison, David and John M. Pitre. (1988). "Computerized Pre-checking of Laboratory Data." The Physics Teacher, 26(3): 156, March.

Heikenkinen, Henry and Gary E. Dunkleberger. (1985). "On Disk With Mastery Learning." The Science Teacher, 52(7): 26-28, October.

Herman, Joan L. and Shari Golan. (c.1992). Effects of Standardized Testing on Teachers and Learning--Another Look. CSE Technical Report 334. National Center for Research on Evaluation, Standards and Student Testing, University of California at Los Angeles.

Huang, Samuel D. and Jane Aloi. (1991). "The Impact of Using Interactive Video in Teaching General Biology." American Biology Teacher, 53(5): 281-284, May.

Hueftle, Stacey J., Steven J. Rakow and Wayne W. Welch. (1983). Images of Science: A Summary of Results from the 1981-82 National Assessment in Science. Science Assessment and Research Project, University of Minnesota, Minneapolis.

Humrich, Eve. (1988). "Sex Differences in the Second IEA Science Study--U.S. Results in an International Context." Paper presented at the annual meeting of the National Association for Research in Science Teaching Lake of the Ozarks, MO, April 10-13.

Jackson, Bryan. (1988). "A Comparison Between Computer-Based and Traditional Assessment Tests, and Their Effects on Pupil Learning and Scoring.", School Science Review, 69(249): 809-815, June.

Jacobson, Willard J., Rodney L. Doran, Edith Y.T. Chang, Eve Humrich and John P. Keeves. (1987). The Second IEA Science Study--U.S.. Second IEA Science Study, Teachers College, Columbia University, New York, NY.

Jacobson, Willard J., Rodney L. Doran, Eve Humrich and Ira B. Kanis. (1988). "International Science Report Card." Teachers, College, Columbia University New York, NY; Washington, DC: National Science Teachers Association.

Kanis, Ira B., Rodney L. Doran and Willard J. Jacobson. (1990). Assessing Science Laboratory Process Skills at the Elementary and Middle/Junior High Levels. Second IEA Science Study, Teachers College, Columbia University, New York, NY.

Keys, Carolyn. (1992). "Evaluating Students' Written Reports for Evidence of Conceptual and Procedural Understandings in Science." Poster session presented at the annual meeting of the National Association for Research in Science Teaching, Boston, MA, March 21-25.

Knight, Carlton W. II. and Gary E. Dunkleberger. (1977). "The Influence of Computer-Managed Self-Paced Instruction on Science Attitudes of Students." Journal of Research in Science Teaching, 14(6): 551-555, November.

Krynowsky, Bernie. (1988). "Problems in Assessing Student Attitude in Science Education: A Partial Solution." Science Education, 72(4): 575-584.

Kulm, Gerald and Shirley M. Malcolm, Eds. (1991). Science Assessment in the Service of Reform. Washington, DC: American Association for the Advancement of Science.

Kumar, David D. (1991). "A Note on 'Hyperequation'." An occasional paper. National Center for Science Teaching and Learning, The Ohio State University, Columbus, September.

Lapointe, Archie E., Janice M. Askew and Nancy A. Mead. (1992). Learning Science. Princeton, NJ: Educational Testing Service.

Lomask, Michal, Larry S. Johnson and Laurin P. Hafner. (1992). "Interactive Videodisc as a Tool for Assessing Science Teachers' Knowledge of Safety Regulations in School Laboratories." Paper presented at the annual meeting of the National Association for Research in Science Teaching, Boston, MA, March 21-25.

Lunetta, Vincent, Avi Hofstein and Geoffrey Giddings. (1981). "Evaluating Science Laboratory Skills." The Science Teacher, 48(1): 22-25, January.

Martens, Mary Lee. (1992). "Effects of State-Mandated Testing on Local Science Programs: Case Study of a Suburban Elementary School." Paper presented at the annual meeting of the National Association for Research in Science Teaching, Boston, MA, March 21-25.

May, Peter M., Kevin Murray and David R. Williams. (1985). "CAPE: A Computer Program to Assist with Practical Assessment." Journal of Chemical Education, 62(4): 310-312, April.

Meng, Elizabeth and Rodney L. Doran. (1990). "What Research Says . . . About Appropriate Methods of Assessment." The Science Teacher, 28(1): 42-45, September.

Mihkelson, Ann et al. (1984). "The Application of Computer Technology in the Assessment of Large Undergraduate Classes." Journal of Chemical Education, 61(7): 631, July.

Milkent, Marlene M. and Wolff-Michael Roth. (1989). "Enhancing Student Achievement Through Computer-Generated Homework." Journal of Research in Science Teaching, 26(7): 567-573, October.

Moore, Richard W. and Frank X. Sutman. (1970). "The Development, Field Test and Validation of an Inventory of Scientific Attitudes." Journal of Research in Science Teaching, 7(2): 85-94, June.

Mullis, Ira V.S. and Lynn B. Jenkins. (1988). The Science Report Card: Elements of Risk and Recovery. Trends and Achievement Based on the 1986 National Assessment. Princeton, NJ: Educational Testing Service.

Munby, Hugh. (1983a). An Investigation Into the Measurement of Attitudes in Science Education. Columbus, OH: SMEAC Information Reference Center, The Ohio State University. (ERIC Document Reproduction Service No. ED 237 347).

Munby, Hugh. (1983b). "Thirty Studies Involving the 'Scientific Attitude Inventory': What Confidence Can We Have in This Instrument?" Journal of Research in Science Teaching, 20(2): 141-162, February.

Munby, Hugh, Richard Kitto and Robert Wilson. (1976). "Validating Constructs in Science Education Research: The Construct 'View of Science'." Science Education, 60(3): 313-321.

Nachmias, Rafi and Marcia C. Linn. (1987). "Evaluations of Science Laboratory Data: The Role of Computer-Presented Information." Journal of Research in Science Teaching, 24(5): 491-506, May.

National Center for Education Statistics. (1991). The 1990 Science Report Card: NAEP'S Assessment of Fourth, Eighth, and Twelfth Graders. Washington, DC: U.S. Department of Education

Omerod, M.D. and D. Duckworth. (1975) Pupils' Attitudes to Science: A Review of Research. Windsor, England: NFER Publishing Company.

Raizen, Senta A., Joan B. Baron, Audrey B. Champagne, Edward Haertel, Ina V.S. Mullis and Jeannie Oakes. (1989). Assessment in Elementary School Science. The National Center for Improving Science Education; The NETWORK, Andover, MA, Washington, DC and The Biological Sciences Curriculum Study, Colorado Springs, CO.

Raizen, Senta A., Joan B. Baron, Audrey B. Champagne, Edward Haertel, In V.S. Mullis and Jeannie Oakes. (c.1991). Assessment in Science Education: The Middle Years. The National Center for Improving Science Education; The NETWORK, Andover, MA, Washington, DC and The Biological Sciences Curriculum Study, Colorado Springs, CO.

Schibeci, R.A. (1984). "Attitudes to Science: An Update." Studies in Science Education, 11: 26-59.

Science Scope, Special Supplement on Assessment. 15(6), March.

Shavelson, Richard J., Neil B. Carey and Noreen M. Webb. (1990). "Indicators of Science Achievement: Options for a Powerful Policy Instrument." Phi Delta Kappan, 71(9): 692-697, May.

Summers, Gerald. (1984). "Testor." Journal of College Science Teaching, 13(5): 356-358, March/April.

Swain, Julian. (1985). "Toward a Framework for Assessment." School Science Review, 67(238): 139-159.

The State Department of Education. (1989). New York State Program Evaluation Test in Science. The University of the State of New York; The State Department of Education, Albany, NY.

Uri, Ganiel and Avi Hofstein, (1982). "Objective and Continuous Assessment of Student Performance in the Physics Laboratory." Science Education, 66(4): 581-591.

Vogel, George. (1985). "A Microcomputer-Based System for Filing Test Questions and Assembling Examinations." Journal of Chemical Education, 62(11): 1024-1027, November.

Waugh, Michael Leonard. (1985). "Effects of Microcomputer- Administered Diagnostic Testing on Immediate and Continuing Science Achievement and Attitudes." Journal of Research in Science Teaching, 22(9): 793-805, December.

Waugh, Michael L. and Dawn Currier. (1986). "Computer-Based Education: What We Need to Know." Journal of Computers in Mathematics and Science Teaching, 5(3): 13-15, Spring.

Willis, John A. (1988). "Learning Outcome Testing: A Statewide Approach to Accountability." Technological Horizons in Education, 16(1): 69-73, August.